NRL Report 9351

Preparing a Sublanguage Grammar

DENNIS PERZANOWSKI AND ELAINE MARSH

Navy Center for Applied Research in Artificial Intelligence
Information Technology Division

October 31, 1991

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 31, 1991	Final 1984-1988

4. TITLE AND SUBTITLE

Preparing a Sublanguage Grammar

5. FUNDING NUMBERS

PE - 62234N
TA - RS34-C74-000

6. AUTHOR(S)

Dennis Perzanowski and Elaine Marsh

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
Washington, DC 20375-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL Report 9351

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
Arlington, VA 22217

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

In this report we discuss the methods used in the preparation of a sublanguage grammar for processing particular sets of Navy messages. We used a set of Casualty Reports or "CASREPS" dealing with the failure of Starting Air Compressors. We also discuss our computational solution for processing a particular grammatical problem in several of the messages, namely so-called "garden-path" sentences. We offer our linguistic motivations for successfully parsing such constructions.

14. SUBJECT TERMS

Message processing
Natural language processing
Natural language understanding
Sublanguage analysis
Syntactic parsing
Text processing

15. NUMBER OF PAGES

30

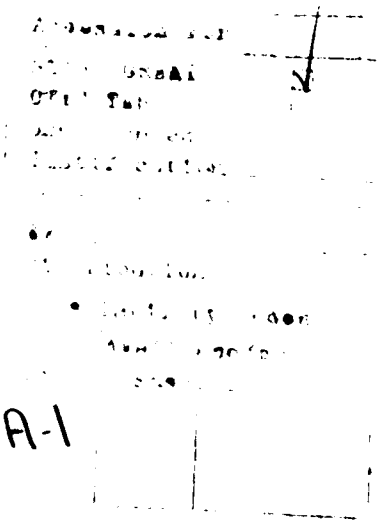
16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18
298-102

CONTENTS



PREPARING A SUBLANGUAGE GRAMMAR

1. INTRODUCTION

The first step in preparing a sublanguage grammar to parse a given set of messages in a particular subdomain, namely Navy Casualty Reports ("CASREPs") of the failure of Starting Air Compressors, was the porting of an English grammar from The Linguistic String Project (hereafter LSP) at New York University.[1] This grammar consists of:

- a set of BNF productions in the **Syntactic Component**,
- a series of LISTs in the **LIST Component** where generalizations are further codified, making the processing of sentences more efficient,
- a set of syntactic-semantic Restrictions in the **Restriction Component** that constrains the productions of the grammar further,
- a set of syntactic transformations and regularizations in the **Transformation Component** that regularizes the various types of sentences parsed into similar structures, and
- a set of Formatting Rules in an **Information Formatting Component** that maps syntactic structures into information structures.[2]

In this report, we will be concerned only with the adaptations made to the syntactic or **BNF Component**, to the **LIST Component**, and to the **Restriction Component**.[3]

Next, to enable the ported grammar to parse sentences from a specific domain, a dictionary was compiled in which the words from a given corpus of sentences were classified into the principal parts of speech and subcategorized for various co-occurrence patterns. Thus, a word like **CONDITION** is classified as a NOUN and a VERB and the principal parts or forms of the word are encoded into a lexicon. Each of these forms, furthermore, is subcategorized for co-occurrence constraints; that is, constructions that may or may not co-occur with the particular form of the word are listed. Non-co-occurrence constraints are listed here to speed parsing. Figure 1 presents a sample lexical item.

```
(NVTV) CONDITION.
.11 = NONHUMAN, NCOUNT1, NAV-STATUS.
.12 = OBJLIST: .3 NOTNOBJ: .1, NAV-REPAIR.
.3  = NSTGO, NTOVO.
.1  = NTIME1.
(TVVEN) CONDITIONED ^.
.14 = OBJLIST: .3, NOTNOBJ: .1, POBJLIST: .4, NAV-REPAIR.
.4  = TOVO, NULLOBJ.
(ING) CONDITIONING ^.
(NTV) CONDITIONS ^.
```

Fig. 1 — A sample lexical entry

Manuscript approved July 8, 1991.

1. We direct the reader to Ref. 1 for a complete description of the porting of the LSP grammar to the Navy subdomain.
2. Some discussion for expository clarity will be offered below; however, the reader is directed to Ref. 2 for a complete discussion.
3. The **Transformation** and **Regularization Components** were stabilized during the porting of the grammar to the Navy domain. No discussion is offered here. Changes made in the **Information Formatting Component** are discussed in [3].

The principal parts, such as noun (N), verb (V), tensed verb (TV), present (ING) and past (VEN) participles, are further codified into canonical forms of (NVTV), (TVVEN), (ING), and (NTV). All morphologically related forms of a word are "uparrowed," a notational convention indicating relatedness, to the base form. This convention allows all forms of a word to share in certain lexical subcategorizations. The various numerical attributes are also notational conventions to allow the parser to easily identify the various attributes that characterize the lexical items. These attributes characterize the kinds of syntactic constructions that co-occur with the various forms of the lexical item, such as nominal objects of transitive verbs. In Fig. 1, transitivity, for example, is specified by **NSTGO** (Noun STrinG Object) as one of the .3 attributes of the **OBJLIST**[4] to (NVTV) **CONDITION**.

Lexical items may also be subcategorized for elements with which they never co-occur. For example, **CONDITION** never takes an object that is subcategorized as an **NTIME1** word. This is indicated in Fig. 1 by the **NOTNOBJ** in line .12 as constrained by **.1 = NTIME1**. This may seem to be a redundant usage of lexical subcategorization, but given the LSP parser,[5] sometimes strange parses can be obtained because negative co-occurrence constraints had not been stipulated. Thus, a simple sentence like ***The Starting Air Compressor failed for a day one month ago*** will yield a strange parse in which ***one month ago*** parses as the direct object of ***failed*** if the object of the verb **FAIL** is not so constrained. Temporal nouns (**NTIME1**) must be prohibited in this environment.

A word can be multiply classified if the word is found in several syntactic environments. Thus, a word like **CONDITION** is both a noun and a verb, indicated by the canonical formula (NVTV). As a verb, for example, several subcategorizations of the word may be permitted for the types of object complements that the verbal sense takes, as indicated by the .3 line of the lexical item in Fig. 1. We will return to this point at some length below, since some very interesting linguistic and computational problems arise as a result.

Finally, the lexical entry of a word might also contain some domain-specific semantic information. For example, the noun **CONDITION** is subcategorized as **NAV-STATUS** on line .11 of Fig. 1 and the verb **CONDITION** is **NAV-REPAIR** in lines .12 and .14. These domain-specific semantic classes, which coincidentally happen to be disparate for the two classifications of the word **CONDITION**, are derived by distributional analysis [4], as are the other classifications and subcategorizations of lexical items. These semantic classes are later used to group lexical items into patterns that are characteristic of the sublanguage under investigation. These latter issues will not concern us here, although some reference to these semantic classes will be made here as they affect the **Restriction Component** of the sublanguage grammar, and introduce issues discussed in Ref. 3.

Briefly, the BNF rules are syntactic productions that expand a single syntactic category, such as **SENTENCE** or **CENTER**, into one or more possible syntactic options. Parts of sentences, therefore, are attached at various points or nodes of a parent string, and these subsequent strings are themselves modified by further expansions until some terminal or final node is obtained. Figure 2 presents a sample of some of the BNFs in the Navy sublanguage grammar.

```
<SENTENCE>  ::= <CENTER> "." .
<CENTER>    ::= <ASSERTION> / <FRAGMENT> .
<FRAGMENT>  ::= <SA> ( <TVO> / <SOBJBESHOW> / <VINGO> / <VENPAS> /
                    ( <NSTG> / <ASTG> / <PN> ) <SA> ).
```

Fig. 2 — Some BNFs in the Navy Sublanguage Grammar

4. The **OBJLIST** refers to the lexical subcategorization of verbs, specifying the classes of **OBJ**ects that can be **LIST**ed as co-occurring with a particular verb. Thus, a verb like **INVESTIGATE** will have a .3 attribute in its **OBJLIST** specifying **NSTGO**. This subcategorization indicates that the verb **INVESTIGATE** is a transitive verb, as in ***Ship investigated the cause of the failure*** where ***cause*** is taken as the **NSTGO** [Noun STrinG Object] of the verb ***investigate.***

5. The LSP parser is a top-down, left-right deterministic parser.

For example, the BNFs in Fig. 2 state that sentences, or **SENTENCE**s, of the sublanguage under investigation can be analyzed syntactically as strings of two elements, namely a syntactic category, here called **CENTER**, and a terminal mark of punctuation, indicated by the period inside the quotation marks. Because the Natural Language Project of the Navy Center for Applied Research in Artificial Intelligence has been involved in text processing, punctuation marks and various other syntactic idiosyncrasies of text are included in our definitions. A grammar attempting to define spoken English, or spoken sublanguages, would, of course, have to account for the distribution of data in a different way. The final punctuation marks in the productions in Fig. 2 are not parts of English strings but are terminations of the BNFs and are used by the parser in expanding and terminating the productions of the grammar.

A complete grammar of English, or a grammar of a different domain [5], might also include the fact that **SENTENCE**s of English or the particular domain under investigation consist of **QUESTION**s and **IMPERATIVE**s requiring the addition of these syntactic categories as well as their corresponding punctuation marks to their respective BNFs. However, these latter types of **SENTENCE**s do not appear in the large body of data investigated.[6] To simplify our discussion, we simply omit their inclusion in the BNFs and do not discuss them here. In decomposing **SENTENCE**s in the CASREP domain, we use the second rule in Fig. 2 which states that **CENTER**s consist of either an **ASSERTION** string or a **FRAGMENT** string, these being the two most common types of **CENTER** strings in this domain. These latter syntactic categories are further decomposed into their constituent strings. BNFs are further constrained in that only one syntactic category is permitted to expand, and the order of options on the right-hand side of the rewrite symbol (::=) indicates the order in which those options are chosen and expanded further by the parser.

Because the LSP parser is a deterministic backtracking parser with limited work space, the ordering of options in the original BNF can oftentimes be crucial in obtaining a good parse. The interaction of lexical subcategorization, parsing algorithm, and type of parser will cause the parser to automatically select the first option encountered that is identical to the specific subcategorization of the word currently being parsed. If the BNF option chosen for expansion is a correct one for that lexical item, but is not the correct one for the structural description of the entire sentence being parsed, the parser may very likely arrive at an incorrect parse, or run out of nodes by either backtracking or garden-pathing. We will discuss some of the problems associated with the latter and related parsing strategies in Section 6 at greater length.

In preparing the Navy grammar, we found it necessary to adapt English BNFs for the following reasons: interaction of lexical subcategorization and the parsing algorithm caused us to reorder existing options in English BNFs; and domain-specific constructions in Navy messages caused us to add options to existing English BNFs or to add new BNFs to the Navy grammar. In the following discussion, we see examples of both types of changes. However, we must first discuss the **Restriction Component**, another component of the grammar where a number of grammatical changes were made.

Restrictions constrain the output of the parser either while the various syntactic categories are expanding by preventing the attachment of inappropriate constructions, or they force the parser to reject a structure for various specified reasons after it has been generated. The parser detaches those constructions and tries other options. These Restrictions, called "D" (disqualify) and "W" (wellformedness) Restrictions, respectively, are ***if/then*** rules. Figure 3 presents examples of the two types of Restrictions.

```
DNAV1 = IN INTRODUCER RE OPTION LNR:
        THERE IS A ':' AHEAD.
WNAV17 = IN NVAR AFTER VING:
        CORE IS NOT VING:NAV-CONN.
```

Fig. 3 — Some Navy Restrictions

6. The entire corpus contains 824 sentences. While all of these sentences have not been processed, investigations of a large part of the corpus over the years have only revealed declarative **SENTENCE**s in this domain.

Basically, **DNAV1** prevents the parser from trying to attach a particular node called **INTRODUCER** unless a colon appears in the sentence. For example, in the sentence ***SITREP 001: SAC failed*** the **INTRODUCER** is the expression ***SITREP 001:***. However, unless the grammar contains a Restriction like **DNAV1**, the parser will always try to attach the **INTRODUCER** node at the beginnings of all sentences. While this is not problematic in the sentence just cited, useless parsing can be avoided in such a **SENTENCE** as ***SAC failed*** where there is no **INTRODUCER**. Furthermore, if the sentence is sufficiently long or complicated syntactically at the beginning, the parser may be garden-pathed simply because a Restriction, such as **DNAV1**, did not keep the gate to that particular garden path closed.

WNAV17, on the other hand, allows the parser to attach a **Noun** node in a particular noun string. Since present participles, or **VING**-constructions, can be used as nouns in English and in sublanguages, the parser must allow the nodes to be constructed, but then after construction, it must check that certain conditions do not hold; otherwise, faulty parses result. Writing **WNAV17** as a "D" (**Disqualify**) condition would be inappropriate to rule out offending constructions, since we might want a gerund or participial noun to be parsed, as in ***Proper maintenance required adequate cleaning of equipment***, where ***cleaning*** is a participial in a gerundive construction. Given **WNAV17**, ***cleaning***, which is not subcategorized as a **NAV-CONN** participle in the lexicon, will parse correctly as a nominal, while other **VING**-constructions that are subcategorized as **NAV-CONN** will not. Without **WNAV17** in the grammar, a sentence like ***SAC failed resulting in shutdown*** might have incorrectly parsed with ***resulting*** as the participial **OBJECT** of the transitive verb ***fail.***

Figure 4 provides an overview of the various grammatical components of the system and how sentences in a corpus are decomposed and analyzed by the grammar. Figure 5 is a more detailed sketch of the grammatical analysis associated with the grammatical component in Fig. 4. Figure 5 provides a flowchart of the various steps that are required in adapting both the various dictionaries and grammars that are used to parse Navy sentences. Figure 5 also indicates how the various components are updated and parsing runs proceed until a good parse is obtained and final forms of a Navy dictionary and grammar are obtained.

In the writing of the sublanguage grammar, we were required to adapt some English Restrictions and to add a number of Navy-specific Restrictions. The English Restrictions were modified for several reasons. As written, several of these Restrictions were too narrow, i.e., the possible syntactic environments were underspecified. We, therefore, had to expand the number of possible options to allow the Navy messages to be parsed based upon the data at hand. A number of Navy-specific Restrictions were also added to the set of Restrictions in the grammar. While the English Restrictions were modified on the basis of Navy data, the syntactic constructions that motivated these changes can be argued to be applicable to English, and are not Navy-specific. The Navy-specific additions are clearly domain-specific and were made either to optimize the parsing of domain-specific constructions or to restrict the occurrences of these constructions.

During the updating procedure, numerous factors and their interactions can influence the parsing; therefore, extensive daily logs were kept. These logs enabled us to retrace our analytical steps and rationale for individual changes. Examples of our extensive log keeping are given in Appendices A and B. Appendix A contains an example of the daily logs kept during the updating procedure as each sentence was analyzed, and Appendix B contains an example from a summary log of grammatical changes. The latter log was maintained to keep all of the grammatical changes together in one place after they had been integrated into the grammar, and to keep older forms of rules, if it became necessary to resurrect an older form of a rule.

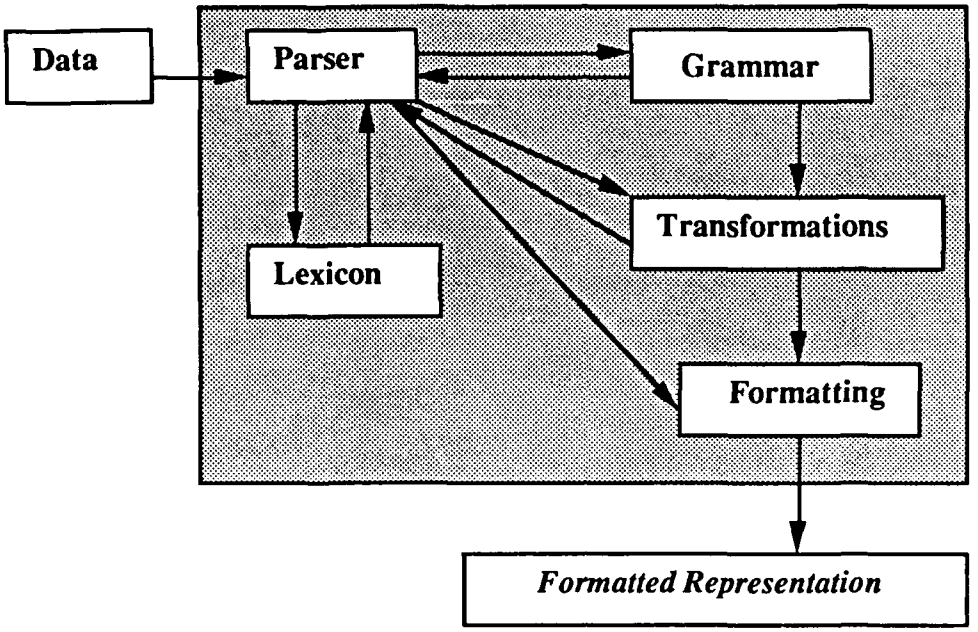


Fig. 4 — Processing sentences through LSP grammar

2. INTERACTIONS OF LEXICAL SUBCATEGORIZATION AND PARSING ALGORITHM

We found that it was necessary in some cases to reorder existing options in English BNFs. Two BNFs were affected, namely **OBJECT** and **CSSTG**, the latter being a mnemonic for subordinate conjunction strings. **OBJECT** is the LSP category that typically expands into object strings in English, such as in the sentence ***This situation presents a hazard*** where ***hazard*** is the object string of the verb. **CSSTG** expands syntactically into various types of subordinate clauses, such as the string ***while the engine started*** in the sentence ***SAC failed while the engine started***. In both cases, reordering of options in these two syntactic categories was required because of the interaction of lexical subcategorization and the order of subcategorized options when a BNF was expanded. In the next section, we will elaborate on the reasons for this particular grammatical change.

Changes in Lexical Subcategorization

In sentence (1), **REMAIN** can be subcategorized lexically so that its **OBJLIST** will permit a participial construction, such as **FULLY ENGAGED**.

(1)[7] [Testb 11.1]: COMPRESSOR WILL NOT REMAIN FULLY ENGAGED CAUSING ERRATIC OPERATION, SURGING AND A HAZARD TO PERSONNEL AND EQUIPMENT.

7. In the following examples, the sentences from the various messages studied are unedited, unless otherwise indicated. These sentences are preceded by a Sentence Identification Number. Thus, in this sample, "Testb" is the name of the batch of sentences the example comes from. "11.1" indicates that the sentence comes from the eleventh message from the Testb batch, and it is the first sentence of that message.

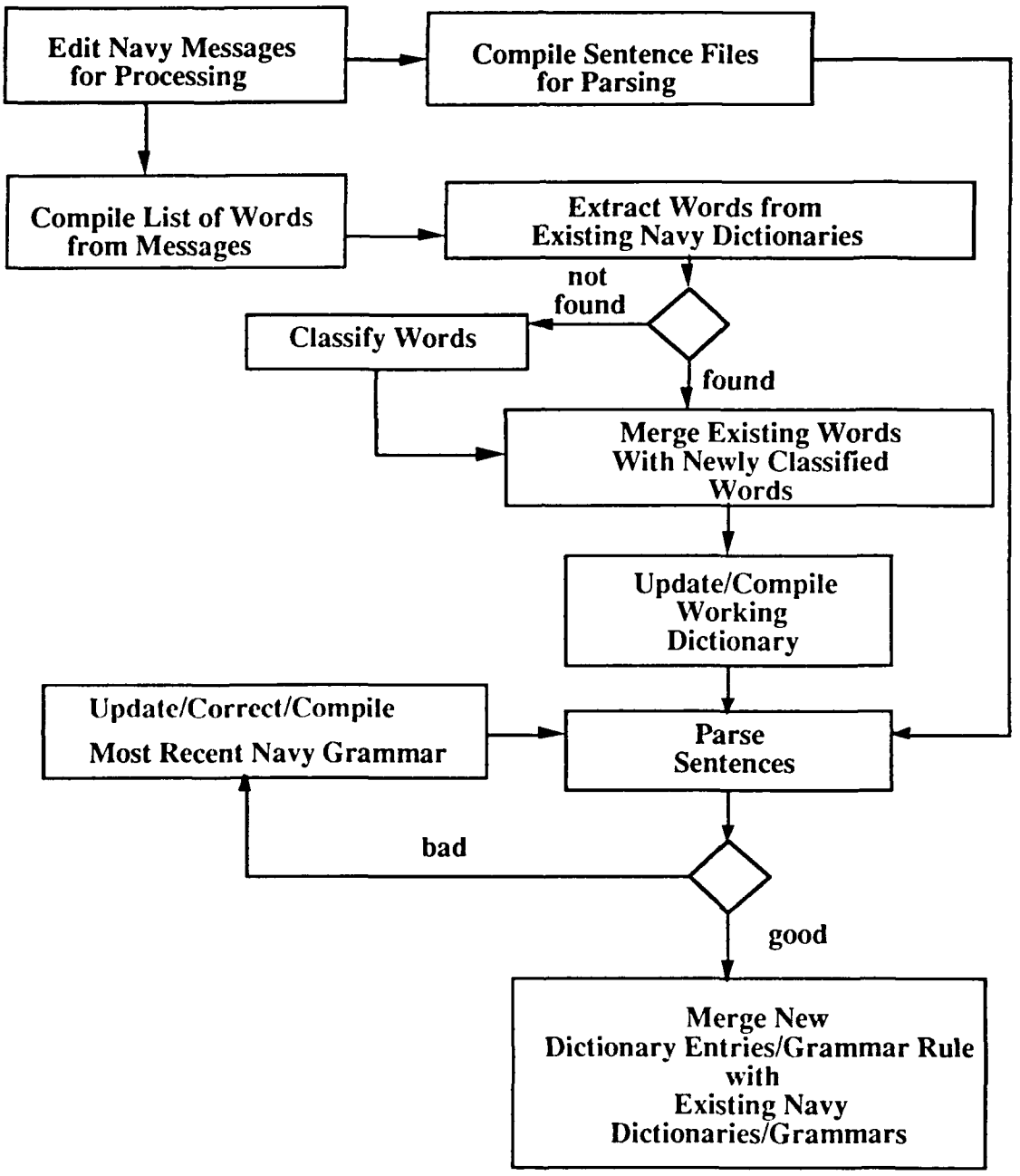


Fig. 5 — Grammatical analysis in preparing a sublanguage grammar

However, **REMAIN** can also be subcategorized in its **OBJLIST** for **NULLOBJ**, as in ***The condition remained*** where the **OBJECT** of ***remained*** is empty or **NULL**, yielding an intransitive reading. If **NULLOBJ** is ordered before **OBJBE** in the BNF definition of **OBJECT**, the parser will automatically select the intransitive option for **REMAIN**, namely the **NULLOBJ** option and will move on to the next node and try to parse the remainder of the sentence accordingly. In (1), it will try to parse **FULLY ENGAGED** as a participial in the sentence. If its structural description is met, as for example as a sentential participial modifier, then the parsing will terminate, having arrived at a successful parse. However, in (1), the latter parse is bad. We, therefore, reordered the **NULLOBJ** and **OBJBE** options in **OBJECT** to force the parser to select the **OBJBE** option before the **NULLOBJ** option in **OBJECT**.

Likewise, reordering of BNF options was required to parse sentences like (2) and (3).

(2) [Srepa 4.2]: SAC WAS SEPARATED FROM SSDG REVEALING O-RING ON FORWARD END OF SPLINE DRIVE SHAFT TOWARDS SSDG TO BE COMPLETELY DESTROYED, ALLOWING SPLINE DRIVE SHAFT TO SLIDE FORWARD, DISENGAGING FROM HUB DAMPER ASSEMBLY.

(3) [Srepa 5.3]: S/F INVESTIGATED AND FOUND CAUSE TO BE DEFECTIVE SAC INPUT DRIVE SHAFT AND HUB DRIVE ADAPTOR.

The verb **REVEAL** in (2) and the verb **FIND** in (3) are commonly subcategorized for **NSTGO** in their **OBJLISTs**, i.e., transitive readings of these verbs are fairly common. Thus, sentences like ***Investigation revealed failure of SAC***, and ***Investigation found the cause of the failure*** are accounted for by the subcategorization of these verbs with **NSTGO** in their respective **OBJLISTs**.

On the other hand, if **NSTGO** precedes the expansion of **NTOBE** in the BNF **OBJECT**, which is the desired verbal complement in these two sentences, both (2) and (3) will be parsed with direct objects ***O-RING*** and ***CAUSE***, respectively. Furthermore, because of the length of (2), with ***O-RING*** as the direct object of the verb, the parser will then try to parse the remainder of the sentence as some kind of sentential modifier. And in (3), the parser will parse **CAUSE** as the direct object of the verb **FIND**. The parser will then try to parse the remaining infinitival construction as a sentential adjunct. These facts, further complicated by the conjunction in (3), will cause the parser to run out of allocated work space, unless an inordinate amount of space is pre-allocated, and terminate with no acceptable parse.

Both **REVEAL** and **FIND** can be subcategorized for **NTOBE** in their **OBJLISTs**, and reordering the **NTOBE** and **NSTGO** options in the Navy definition of **OBJECT** produces correct parses for sentences (2) and (3) and their like. Therefore, if a verb is classified for both direct objects and for embedded infinitival clauses with overt subjects, the **N** of **NTOBE**, the noun will be parsed correctly. These results are consequences of our work on the interaction of the lexical subcategorization of verbs, the ordering of options in BNF definitions, and the backtracking that is available in the parser when requisite structural description of elements in the string are not met. In the next section, we discuss the reordering of syntactic options in the expansion of **CSSTG**.

While these reorderings in BNFs were a satisfactory solution to handle the parsing problem, it raises the issue of whether or not such reorderings will adequately handle similar constructions not yet encountered. The question to be answered, therefore, is: do all verbs subcategorized for **NULLOBJ** and **OBJBE** (as in (1)) or for **NTOBE** and **NSTGO** (as in (2)) act similarly in sentential environments? In other words, if a verb is doubly subcategorized for **NULLOBJ** and **OBJBE**, is it correct to assume that the **OBJBE** subcategorization should be invariably processed first? Similarly, verbs doubly subcategorized for **NTOBE** and **NSTGO** will be processed. Our grammatical change seems to be making this claim, but it is subject to further empirical verification, which was not undertaken during this study.

Reordering Options for Subordinate Clauses

Like the changes made in the options of **OBJECT**, the **SUB3**[8] option in **CSSTG** was reordered because of lexical subcategorization and interaction with the parsing algorithm. In (4), the subordinating conjunction **WHILE** is multiply classified in the lexicon as a subordinating conjunction. These classifications (**CS1**, **CS3**, among others) are based on its distribution in subordinating clauses such as those in (4a-b).

(4) a. WHILE SAC WAS DISENGAGED, CASUALTY OCCURRED. (WHILE = CS1: pre-ASSERTION)
b. WHILE STARTING GAS TURBINE, NR 2 SAC EXPERIENCED LOSS OF L/O PRESSURE. (WHILE = CS3: pre-progressive participle)

8. While the names of grammatical categories or nodes is chosen for strictly mnemonic reasons, their syntactic behavior is determined by "distributional analysis" [4]; therefore, their syntactic identity and behavior is empirically derived.

By ordering **SUB3** (::= <CS3> <VINGO>) before **SUB1** (::= <CS1> <ASSERTION>), a more efficient and speedier parse was obtained in parsing sentences like (4b).

The string **STARTING GAS TURBINE** in (4b) is syntactically ambiguous. It can be parsed as a nominal string with the progressive participial as a left adjectival modifier of the nominal expression **GAS TURBINE**. If **SUB1** is expanded first, the parser attaches **ASSERTION**; next, the **SUBJECT** of the **ASSERTION** is expanded into the erroneous nominal just cited. With the **SUBJECT** completed, nothing is left in (4b) to satisfy the structural description of **ASSERTION**; consequently, the parser will have to detach several nodes and back up to attach **SUB3**, a subsequent option in **CSSTG**. With **SUB3** ordered before **SUB1** in **CSSTG**, unnecessary backtracking is avoided.

An alternative solution would have been to write a Restriction sensitive to the presence of the comma at the end of the introductory subordinate clause. Clearly, its presence aids the reader in parsing the initial string correctly; however, we do not believe that we can prescriptively guarantee the presence of a comma in this environment. Thus, a grammatical rule that forces the introductory participial reading based on the presence of an upcoming mark of punctuation would fail if the writer of the message had forgotten to include the comma. Therefore, a reordering solution seems to be the more justified solution.

Finally, to reduce parsing time by eliminating the number of superfluous options available in parsing subordinate clauses, we removed **SUB9** from the expansion of **CSSTG**. **SUB9** expands into a construction like the introductory clause found in (5).

(5) SHOULD YOU FIND THE LUBE OIL PRESSURE LOW, YOU MAY HAVE TO REPLACE THE SAC.

In the corpus surveyed for this grammar, we have no instances of modal auxiliaries in subordinate clauses, such as the use of **SHOULD** in (5). We, therefore, eliminated the **SUB9** option of **CSSTG** from the Navy grammar.

3. EFFECT OF DOMAIN-SPECIFIC CONSTRUCTIONS

Domain-specific constructions in Navy messages have caused us to modify English BNFs in two ways. In some cases we rewrote the English BNF; in others, we added new BNFs based on the Navy CASREP data.

Rewriting of BNF Options

Rewriting Options for Subordinate Clause Strings

Originally, **SUB6** and **SUB7** were "rare" options in the English grammar. This was signified by the notational convention of prefixing a hyphen before these options in the expansion of **CSSTG**. This notational device triggered the **Rare Mechanism** to operate during the parsing process. Instead of expanding a particular option, the option was skipped if rare and was only expanded if the **Rare Switch** was turned on, or given a value of **True** during the parsing process. Based on the sentences in the CASREP data, we found it necessary to derarify the expansions of **SUB6** and **SUB7** in **CSSTG**.

An Example of Subordinate Clauses

SUB6 expands to a **CS6** subordinating conjunction, such as **WITH**, followed by an **SOBJBE** string. In a set of messages that was analyzed prior to our analysis of Testa, Testb, and Srepa, the structural description for a **SUB6** string was found in sentence (6)

(6) [Qreps 1.5]: WITH CU-2007 ANT COUPLER INOP, CAPABILITIES LOST ARE AS FOLLOWS: NO VLF BROADCAST, NO MONITORING OF THE 500 KHZ EMERGENCY BAND, AND NO SHIP'S ENTERTAINMENT.

In (6), the introductory clause **WITH CU-2007 ANT COUPLER INOP** can be analyzed as a **SUB6** string. **WITH** is subcategorized as a **CS6** subordinating conjunction. It is followed by the fragmented **SOBJBE** clause, characteristic of **SUB6** strings. The BNF **SOBJBE** expands into **SUBJECT**, followed by the predicate adjective

INOPER[ATIVE].[9] Since Qreps 1.5 exhibited a **SUB6** construction and the **Rare Switch** was broken, the option had to be derarified.

Another Example in Subordinate Clauses

SUB7 was derarified to account for the sentences in (7).

(7) a. [Testa 4.1]: WHILE DIESEL WAS OPERATING WITH SAC DISENGAGED, SAC LO ALARM SOUNDED.

b. [Srepa 1.1]: NR 4 SSDG STARTED WITH SAC DISENGAGED AND LOW LUBE OIL PRESSURE ALARM INDICATED.

c. [Srepa 3.1]: UNABLE TO MAINTAIN MINIMUM OIL PRESSURE WITH UNIT NOT ENGAGED.

d. [Srepa 8.3]: REFILLED SAC WITH OIL AND TEST RAN DIESEL WITH SAC DISENGAGED.

The sentences in (7) contain **SUB7** strings, beginning with the multiply subcategorized word **WITH**, which is both a **SUB6** (cf. (6) and its discussion) and a **SUB7** subordinating conjunction. In (7), the clauses **WITH SAC DISENGAGED** and **WITH UNIT NOT ENGAGED** are **SUB7** strings in **CSSTG. SUB7** strings consist of **SUB7** conjunctions followed by **SVEN** complements, which are analyzed as **SUBJECT**s followed by their participial (**VEN**) complements. Like (6), the sentences of (7) could have been parsed with a derarified **SUB7** option in **CSSTG** had this switch been working.

An Example in Navy-specific Dates

So far, we have discussed changes made to a grammar to account for data from a particular subdomain. However, all of the changes that we discussed are changes that would, no doubt, be required for other English domains as well. Furthermore, several of the changes were made because of the particular parsing algorithm used by the LSP parser, or because of certain problems associated with the parser itself. We will now address the specific changes that were made to the grammar that were prompted by very specific Navy constructions that were discovered in the data.

Navy messages exhibit a somewhat complex string of words, numbers, and letters to express dates as seen in the sentences of (8).

(8) a. [Testb 34.4]: SITREP 001, 120010 Z SEP 81: INVESTIGATION BY TODD REVEALED SAC SPLINE INPUT DRIVE SHAFT DISCONNECTED FROM DIESEL HUB.

b. [Srepa 2.4]: TESTED SAT[ISFACTORY] ON 25 FEB.

c. [Srepa 5.5]: NR 4 SSDG IS EXPECTED TO BE OPERATIONAL BY 1200 2 MAR WHICH WILL ALLOW SHIP TO GET UNDERWAY BUT WILL HAVE NO BACKUP START CAPABILITY.

In (8a), the phrase **SITREP 001** is not part of a Navy date expression. It is an introductory header indicating a **SIT[UATION] REP[ORT]** that is parsed as an **INTRODUCER** (cf. Fig. 3 and the discussion that follows), followed by the number of the updated report. The intermediate comma will not concern us here, its purpose being to separate the **SITREP** header from the date-time string. LSP simply inserts the comma as punctuation.

The complex Navy date **120010 Z SEP 81** follows the **INTRODUCER** in (8a). In this fairly common Navy date-time phrase, all of the elements of military date-time expressions, expressed in "Zulu time," are present. Reading from the left, the first two digits of the complex numeral **120010** indicate the date, i.e., **12.** These are followed by the time, expressed in terms of the twenty-four-hour clock. **0010** indicates **ten minutes after midnight.** The **Z** is the identifier for "Zulu time." This is followed by the month **SEP**, which is usually an abbreviation, and the year, **1981**, abbreviated to the last two digits.

9.Documentation for the Qreps run does not specify why **SUB6** was derarified in parsing Qreps 1.5; however, our research with the **Rare Switch** later indicated that it must have been broken during the porting of the operating system to the Navy domain. By oversight, no doubt, it was never fixed. However, had it been fixed, the **SUB6** option could have remained rare, the switch turned on, and the parser would have expanded this option. Qreps 1.5 could then have been accounted for. (Similarly, **SUB7** could have been handled.)

Sentences (8b and c) also express date-time, but in far less complicated ways. Sentence (8b) merely indicates the date and month, while (8c) cites the time, assumed to be expressed in terms of the twenty-four-hour clock, the date, and the month. In order for LSP to parse the variations in date-time expressions found in Navy messages, it was necessary to introduce an optional **Z** element, for date-time strings expressed in terms of "Zulu time." **Z** was included as an optional element, since it does not always appear in Navy date-time strings, as the sentences in (8) indicate. Also, as (8b and c) indicate, the terminal element in the date time expression need not be a number, or **Q** in terms of LSP notational conventions, but the terminal element can be a noun, or **N**. Therefore, a final optional **Q** was included in the Navy date-time string **DAYYEAR**. By including the optional elements in **DAYYEAR** as stipulated, the Navy sublanguage grammar in LSP was capable of parsing the variety of date-time expressions found in Navy CASREP messages.

Adverbial/Adjectival Modifications in Compound Navy Nominals

Adaptations of the LSP English grammar were also made in two types of adverbial and adjectival expressions. These changes were made in the adverbial and adjectival modification in compound Navy nominals. One change was made in the adverbial modification of compounded clauses. Compounding in nominals is highly productive in English; we found that the CASREP domain also exhibits a high proportion of compound nominals. No statistical figures exist at this time, but it is our impression that in the 824 sentences surveyed for this study, nominal compounding is a highly productive rule in this particular sublanguage. Compounding perhaps arises from compacting as much information in as few words as possible in a message. Therefore, the BNFs and Restrictions dealing with nominal compounding were worked on quite extensively.

Left-branching Adverbial Modifiers in Nominals

Two modifications were made to adverbial elements found in compound Navy nominals, such as those in (9).

(9) a. [Testb 32.3]: THIS SITUATION PRESENTS POTENTIAL OVER TEMP HAZARD TO LM2500 AND FURTHER DEGRADATION OF MOBILITY.
b. [Testa 1.1]: STARTING AIR REGULATING VALVE FAILED.
c. [Testa 6.1]: UNABLE TO MAINTAIN LUBE OIL PRESSURE TO STARTING AIR COMPRESSOR.
d. [Testb 8.1]: LOSS OF ONE OF TWO STARTING AIR COMPRESSORS.
e. [Testb 14.2]: STARTING AIR COMPRESSOR ENGAGED FOR APPROX TWO MINUTES WHEN LUBE OIL PRESSURE DROPPED BELOW 65 PSI [POUNDS PER SQUARE INCH] ALARM SETTING.

In the following discussion, we will look at such Navy compound nominals as **POTENTIAL OVER TEMP HAZARD**, **STARTING AIR REGULATING VALVE**, and **STARTING AIR COMPRESSOR.**

Adverbs that Modify Adjectives — The nominal POTENTIAL OVER TEMP HAZARD in (9a) parses with the adjective POTENTIAL in an expected adjectival slot, namely an APOS node to the left of some host noun. The remainder of this nominal, however, was not parsable as a modifier of the host noun HAZARD. There simply were no available expansions in the grammar. We, therefore, added an adverbial node to the left modifier of a compound-noun, namely LCDN.[10] By incorporating a BNF for LCDN at a sufficiently low level of expansion in the grammar (see Appendix C), we were able to obtain the correct modification so that adverbs in such sentences as (9a) have the adverb OVER modifying TEMP and not HAZARD.

These constructions are typically hyphenated in everyday English, but we do not find their counterparts as hyphenated Navy expressions. We, therefore, had to account for them syntactically by adding an option to the expansion of one of the BNFs, rather than initiating any lexical changes in the words.

More will be said about the parsing of compound Navy nominals in a later section. We will now briefly direct our attention to another type of internal modification in compound Navy nominals.

---

10. **LCDN** is a mnemonic for the Left modifier of a CompounD Noun. Such internal branching of left modifiers is fairly common in ordinary English, as we see in such expressions as *low-life person*, *fail-safe button*, and *off-line storage*.

Present Participles as Adjectives — In (9b-e), the compound nominals **STARTING AIR REGULATING VALVE** and **STARTING AIR COMPRESSOR(S)** exhibit adjectives formed from the present participle of the verb **START**. To process these constructions, it was necessary to derarify the **VING** option in **LCDN**, which is, as noted above, the node in the left modifier of compound nominals that permits internal branching of modifiers in those nominals. Because we noted several instances of this particular kind of adjectival modification in compound nominals, we decided that derarification of this option was totally justified. Such expressions as **STARTING AIR REGULATING VALVE** are analyzed with the host noun **VALVE** modified by the complex nominal modifier **STARTING AIR REGULATING**; the more deeply embedded host noun **AIR** is to the left of the ultimate host noun, and the participles act as left and right modifiers, respectively, of the embedded host noun **AIR**.

Parsing such compound nominals is crucial for the subsequent analysis by the next component, the TExt Reduction SystEm (**TERSE**) [3]. **TERSE** is responsible for analyzing the parsed and formatted text, checking to see which pieces of equipment are being referred to in one of the knowledge bases of that component, and then producing a correct analysis of the text based on user's needs. If the grammar does not parse these nominals correctly, the specific piece of equipment being referred to will not be identified in the knowledge bases, and incorrect or no analysis will be obtained by **TERSE**. Therefore, correct syntactic analysis is crucial at this point to parse and correctly identify the various host nouns and their modifiers, no matter how complex or deeply embedded these categories might be in their respective compounded constructions.

Adverbials in Conjoined Clauses

The second modification made to accommodate adverbial expressions in the Navy grammar was to allow an adverbial expression to appear in conjoined clauses, as in (10).

(10) [Testb 13.1.b]: OIL PRESSURE DROPPED TO 72 PSI, THEN INCREASED TO 90 PSI, AND THEN FAILED WHILE STARTING GAS TURBINE.

While we edited in the commas, the adverbial addition for conjoined clauses is still motivated. The adverbial **THEN** occurs in each of the conjuncts in (10). Since we needed a node in conjoined strings that would allow for the adverbial in those positions, we included **SACONJ**, a node that already existed elsewhere in the English grammar in the expansion of **COMMASTG** in the Navy grammar.

Verbal Modifications

It was necessary to modify the expansion of the noun string **NVAR** to incorporate what would normally be considered "deverbal nouns" if the morphological rules of English had applied. It was also necessary to restrict an existing expansion of **NVAR**, namely the **VING** option. In another case, it was necessary to add to the expansion of the types of fragments found in the Navy grammar.

Deverbal Nouns

The rules[11] that form so-called "deverbal nouns" in English are quite productive. Deverbal nouns are nouns formed from verbs. Thus, for example, if the morphological rules for their production are applied to the verbs in (11), the nominal counterparts in (12) are formed.

(11) a. accompany
b. inspect
c. open
d. signal
e. suspect

11. Cf. Ref. 7.

(12) a. accompaniment
b. inspection
c. opening[12]
d. signal
e. suspect

Suffixation of a morphological affix occurs in (11a-b) to obtain (12a-b). In (12d), no phonological changes distinguish the nominal and verb forms of the word, such as the shifting of word stress. In (12e), on the other hand, stress shifts from the second syllable of the verbal form in (11e) to the first syllable of the nominal in (12e). However, in the Navy CASREP data, we have come across what we believe to be the nominalization of verbs without suffixation or accompanying phonological change,[13] even when the nominalized forms of the words already exist in Standard English.

Consider the sentence in (13).

(13) [Testb 29.1]: FCT OPEN AND INSPECT REVEALED BEARING MATERIAL ON BOTTOM OF STRAINER.

We claim that (13) exhibits a compound subject which is **FCT OPEN AND INSPECT.** The left modifier **FCT** is, we believe, some sort of Navy organization. Sentence (13) is to be interpreted as **FCT['S] OPEN[ING] AND INSPECT[ION] REVEALED BEARING MATERIAL ON BOTTOM OF STRAINER.** While the nominalized forms for the verbs **OPEN** and **INSPECT** already exist in Standard English, we believe that this data exhibits an instance of an alternative subdomain nominalization.

To process these constructions, it was necessary to allow the noun string in LSP, namely **NVAR**, to expand not only to the normal terminal **N**, a lexical **noun**, but also to **V**. While the categorical change exhibited here by **NVAR** becoming either **N** or **V** is empirically unjustifiable, given conditions of Boolean analyzability on grammatical rules [6], this change was maintained because of the various interactions of lexical classification and requirements on expansion of grammatical categories in BNFs.This expansion, furthermore, has to be highly constrained (cf. discussion of changes in the **Restriction Component** below); otherwise, numerous bad parses will be generated when **VERB**s of sentences are mistakenly parsed as **SUBJECT**s.[14]

Infinitival Fragments

To process the wide variety of sentence fragments that Navy CASREP messages exhibit, it was necessary to increase the number of types of fragments in the Navy grammar. **SOBJBESHOW** is one of the more productive expansions of **FRAGMENT** in the Navy grammar. A large majority of sentences in the CASREP corpus exist as fragments. The **SOBJBESHOW** type, namely one in which the **SUBJECT** and a missing copula (i.e., linking verb) **[BE]** are followed by one of several complements, is perhaps the most productive. One complement that was lacking in **SOBJBESHOW** was the **TOVO** type, as seen in (14).

(14) [Testb 34.6]: TODD LA TO REPLACE WORN HUB ASSEMBLY AND SPLINE SHAFT.

We interpreted (14) as an **SOBJBESHOW** fragment with the subject **TODD LA** followed by a **TOVO** complement. This expansion increases even further the number of possible types of fragments in a message processing domain that already exhibits a large number of fragmented sentences in its corpus.

---

12. For simplicity, we are assuming that gerundives are part of the deverbal morphology of English. Cf. Ref. 8.

13. The latter point cannot be proven, since the messages that we have analyzed are not acoustic messages but text messages.

14. Incorporating a rare expansion of **V** in **NVAR** is motivated, given the singular example from the data studied. However, because the **Rare Switch** was broken, we incorporated a normal expansion of **V** in **NVAR** and constrained its occurrences through Restrictions. We further believe that even if the expansion of **V** in **NVAR** were a rare option, we would still want to constrain its occurrence through Restrictions when the **Rare Switch** is turned on.

New BNFs Added to the Navy Grammar

To process the Navy CASREP data, it was necessary to add several new productions to the Navy grammar. These additions were largely necessitated by the complex structure of Navy compound nominals. Several rules were needed, therefore, in the left- and right-hand modifier positions for host nouns, and it was necessary to rewrite the expansion of **SENTENCE** for one rather frequent construction.

Further Modification of Compound Navy Nominals

Left-hand Modification of Compound Navy Nominals

LSP parses two basic types of quantified expressions, such as those in (15).

(15) a. [Testa 31.1]: LOSS OF SECOND OF TWO INSTALLED SAC'S.
b. [Testb 8.1]: LOSS OF ONE OF TWO STARTING AIR COMPRESSORS.
c. [Testb 32.1]: LOSS OF 50 PERCENT OF START AIR CAPABILITY.
d. [Testb 13.1]: OIL PRESSURE DROPPED TO 72 PSI [PRESSURE PER SQUARE INCH], THEN INCREASED TO 90 PSI, AND THEN FAILED WHILE STARTING GAS TURBINE.

The first type of quantified expression can be seen in (15). Sentences (15a-b) exhibit the rather common usage of cardinal and ordinal numbers in quantifying objects in the subdomain, such as **STARTING AIR COMPRESSORS**, and quantified expressions can be used as measurements of properties in the real world, as in (15c-d). The kinds of quantification just cited, however, were treated uniformly in both English and Navy grammars, the rules handling these types of quantification being robust enough to handle the data.

On the other hand, Navy CASREP messages exhibit another type of quantification. In the Navy messages, parts are frequently named by means of numerical expressions. Consider (16) where parts of equipment are referred to by a numeral in their names.

(16) a. [Testa 21.1]: DURING MONITORING OF 1A GRM, NR 4 SAC OIL PRESSURE DROPPED BELOW ALARM POINT OF 65 PSIG [POUNDS PER SQUARE INCH, GAUGE].
b. [Testb 19.1]: REDUCED CAPABILITY OF NR 4 SAC RESTRICTS SHIPS OPERATION.
c. [Srepa 1.1]: NR 4 SSDG STARTED WITH SAC DISENGAGED AND LOW LUBE OIL PRESSURE ALARM INDICATED.
d. [Srepa 5.6]: SITREP 001: SSDG NR 4 SLIPRINGS CORRECTED.

The sentence of particular interest for us in a Navy subdomain is (16d). The string **NR 4** is the numerical identifier of a part. It follows the part **SSDG** that it is identifying, and is itself embedded to the left of a host noun.

To handle the continued parsing of common numerical expressions as well as the unique expressions found in Navy CASREP messages, we were required to add a left-branching structure **LNR1** inside of the noun phrase modifier **NNN**, itself a left-hand modifier of a host noun (Cf. Section 8). By doing so, we were able to parse the name of a piece of equipment followed by its numerical name, which was then followed by a piece of equipment, the host of the entire construction. These additions also permitted the parsing of complex Navy compound nominals as in (17).

(17) [Testb 36.3]: AFTER THE MAINTENANCE WAS ACCOMPLISHED, OPERATIONAL TESTS REVEALED LOW LUBE OIL PRESSURE (65 PSI WHICH IS LOW LUBE OIL ALARM SET POINT) BEFORE THE REQUIRED THREE MINUTE SAC ENGAGED TIME LIMIT HAD RUN OUT.

Sentence (17) includes several compound nominals, but the most complex one is **THE REQUIRED THREE MINUTE SAC ENGAGED TIME LIMIT**. The complexity lies in its multiple nesting and left-hand branching of modified structures in the left-hand modifier of the host noun **LIMIT**.

Working leftward in this nominal, the first nested structure is **SAC ENGAGED TIME**, which modifies **LIMIT**. Internally, **SAC** and **ENGAGED** share constituency, and modify **TIME** as a type of measurement. Con-

tinuing leftward, **THREE MINUTE** modifies **TIME** with **REQUIRED** modifying the constituent **THREE MINUTE**. The brackets in (18) indicate the scope of each of the modifiers and the way in which the Navy sublanguage grammar was tuned to parse left-handed, nested modifications in compound Navy nominals.

(18) [THE [[REQUIRED [THREE MINUTE [[SAC ENGAGED] TIME]]]] LIMIT]

The additions of **LNR1**, **LN1**, and **RN1** in **NNN** allowed the parser to capture and characterize the variety of complex Navy compound nominals.

Left-hand Modification of Navy Nominals by Short Clauses

The rich internal structure of Navy nominals is made more complex by the appearance of "short clauses" as left-hand modifiers to host nouns. By short clauses, we mean embedded propositions lacking overt subjects and having tenseless verbs. For example, English exhibits these type of short clauses in such expressions as **an easy-to-please person** and **a difficult-to-read book**. These expressions are usually hyphenated in standard English text and could be handled as lexical items if such were the case in a subdomain. However, in the texts that we saw, short clauses were not indicated orthographically. We, therefore, had to parse them syntactically, which ultimately seems to be the better solution for purposes of later interpretation.

This latter requirement is not an ad hoc conclusion based upon the lack of punctuation in these cases; on the contrary, parsing these expressions syntactically is well-motivated. Short clauses exhibit an underlying argument structure, and if we are parsing these messages as a first step toward extracting information from text, then parsing these constructions syntactically is a reasonable step to take. As ported, LSP did not have the mechanism to parse such short clauses. Our work on fine tuning the grammar for this particular Navy subdomain produced a grammar capable of handling these constructions in sentences like (19).

(19) a. [Testb 16.1]: DURING NORMAL START CYCLE OF 1A GAS TURBINE, APPROX 90 SEC AFTER CLUTCH ENGAGEMENT, LOW LUBE OIL AND FAIL TO ENGAGE ALARMS WERE RECEIVED ON THE ACC.

b. [Srepa 4.1]: RECEIVED LOW LUBE OIL PRESSURE AND FAIL TO ENGAGE ALARMS WHEN ATTEMPTING TO ENGAGE NR 3 SAC FOR START OF GAS TURBINE ENGINE.

To parse the short clause **FAIL TO ENGAGE** as a constituent modifying **ALARMS** in (19), we added a verb phrase constituent in the left-hand verbal modifier **VPOS** of host nouns.

Right-hand Modification of Navy Nominals

Common right-hand modifiers of nouns in Standard English are prepositional phrases as in (20a), and appositive constructions, as in (20b).

(20) a. THE OPERATION OF THE SAC FAILED.

b. [Testa 28.2]: BLADES ARE BENT AND CHIPS, 1/4 INCH DEEP, ARE VISIBLE ON LEADING EDGE.

Appositives, as in (20b), are handled in a very standard way, so nothing more need be said about them here. In (20a), the prepositional phrase **OF THE SAC** is on the right of the host noun **OPERATION**. However, we noticed in the CASREP data that in some instances prepositions were omitted, as in (21).

(21) [Testb 2.1]: LOSS OF LUBE OIL PRESSURE DURING OPERATION NR. 2 SSDG.

In (21), the preposition **OF** has been zeroed in the larger prepositional phrase **DURING OPERATION [OF] NR. 2 SSDG**. To process this construction in the Navy grammar, it was necessary to add a syntactic category **PARG** (a Preposition-less **ARG**ument) on the right-hand side of a host noun. **PARG** expands the syntactic category of **RNP**, which also expands to **PN** for the complementary prepositional phrases that contain a preposition. **PARG** is also con-

strained severely in the **Restriction Component** based on the subdomain characteristics of the host noun and other linguistically motivated arguments discussed below.

SENTENCE Expansion

The first expansion in the LSP grammar was the expansion of **SENTENCE**. It uniformly expanded into an introductory element followed by a **CENTER** or the major predicate-argument structure of the sentence and terminated in an endmark. The introductory element or **INTRODUCER** expanded into one of the coordinating conjunctions **AND, OR,** and **BUT**. Although conjoining of elements, such as phrases and clauses, was and is handled by a **Conjunction Mechanism** elsewhere in the grammar, **SENTENCE** expanded as it did to permit text containing sentence fragments introduced by one of the coordinating conjunctions. In other words, LSP was capable of parsing a conjoined fragment when that fragment existed in isolation, as in (22).

(22) a. AND PENICILLIN WAS ADMINISTERED.
b. BUT PATIENT DIED.

In the Navy CASREP data, we did not find sentences introduced as they are in (22). On the other hand, we found an introductory element usually consisting of a quantified noun phrase followed by a colon, as in (23).

(23) a. [Testb 7.7]: SITREP 002: DRIVE SHAFT FOR SAC WAS MANUFACTURED LOCALLY.
b. [Testb 34.4]: SITREP 001, 120010 Z SEP 81: INVESTIGATION BY TODD REVEALED SAC SPLINE INPUT DRIVE SHAFT DISCONNECTED FROM DIESEL HUB.
c. [Srepa 9.8]: SITREP 003: REMOVED OLD SAC.

To process the sentences of (23) and sentences like them in the Navy CASREP messages, we rewrote the BNF definition of **INTRODUCER**, constrained it, and altered the expansion of **SENTENCE**. We created an intermediate constituent, **OLD-SENTENCE**, because of interactions between introductory elements, noun phrases, and the category **SENTENCE** during the operation of the **Conjunction Mechanism**. With the intermediate category in the Navy grammar, incorrect conjoinings of introductory elements and nominal subjects of **SENTENCE**s were ruled out automatically.

4. ALTERATIONS TO ENGLISH RESTRICTIONS

As mentioned in Section 1, writing a sublanguage grammar involved work on the **Restriction Component**. In general, work on this component of the grammar consisted mainly of two types of alterations: loosening existing Restrictions in the English portion of the grammar to allow certain Navy constructions to be parsed and adding Navy-specific Restrictions to constrain the parsing of these constructions, along with making other changes in the grammar. The latter work was specifically in the area of optimizing the performance of Navy-specific Restrictions to make the parsing process more efficient for some of the new constructions added to the sublanguage grammar. Secondly, Navy-specific Restrictions were added and/or altered to the grammar to constrain the output as a result of having altered the BNFs in the sublanguage grammar. This report does not present all of the Restrictions that were added and/or altered to the grammar; rather, discussion is limited to the general types of changes that our linguistic modifications took because we can generalize and group them according to the kinds of changes made. A brief example of each type of change to the **Restriction Component** is also provided.

Loosening of Restrictions

Adding or altering Restrictions to the **Restriction Component** of the grammar can be caused by several factors, one of which is the writing of new BNFs. Obviously, if a new BNF is added to the Navy sublanguage grammar, then a new Restriction may be needed or an old one may need to be modified. The new rule may cause the grammar to either over- or undergenerate. As a result, further fine tuning, usually in the manipulations of grammatical Restrictions, is required. Also, additional data can require rewriting of the grammar. In several instances, for example, as in (24), new data was presented and an English Restriction had to be altered to allow these and other sentences like them to parse.

(24) a. [Srepa 10.3]: ASSISTANCE REQUESTED TO REMOVE SCATTER SHIELD WHEN NEW SAC RCVD.

b. [Srepa 1.1a]: NR 4 SSDG STARTED WITH SAC DISENGAGED.

c. [Srepa 3.1]: UNABLE TO MAINTAIN MINIMUM OIL PRESSURE WITH UNIT NOT ENGAGED.

In the medical domain that produced the original LSP grammar, sentence fragments existed and required the production of appropriate BNFs and Restrictions. Data like (24), however, did not exist. Therefore, the BNFs producing **FRAGMENT**s were altered so that such sentences could be generated. As is frequently the case in writing grammatical rules, the writer(s) writes the rules to capture the specific data prompting the writing of the rule and to capture as many other cases as possible without losing efficiency of parsing. Thus, for example, if the rule writer knows that adjectival fragments such as (24c) exist in English, that writer may incorporate such a BNF in the general grammar of English. However, such facts may go unnoticed when the Restrictions are being written.

The **Restriction Component** plays a paramount part in the creation of a sublanguage grammar. Frequently, the original grammar is comprehensive, as was the case with the LSP grammar that was ported to this Navy domain. The original writer(s) of the grammar, wanting to be as comprehensive as possible in writing a broad coverage grammar, may have written BNFs and Restrictions to produce such sentences as (24a-b). However, the writers of the Restrictions for that grammar may not have encountered the wide variety of sentences possible in the data and, therefore, did not write an appropriate Restriction. If an occurrence does not exist in the data under investigation, those writers have no need to constrain the rules to allow for the efficient parsing of such sentences. Therefore, rules governing such output are not written at that time. Given a comprehensive English grammar such as the LSP grammar, a great deal of sublanguage work is devoted to the writing, altering, and refining of Restrictions. Certain constraints may already be included to restrict the occurrences of related constructions; therefore, the interaction of BNF rules and existing Restrictions needs to be observed for adverse consequences. In the case of (24), for example, BNFs had been written to produce such kinds of **FRAGMENT**s. However, a specific Restriction, **DPOS4C**, had to be rewritten to loosen its application and thereby allow all of the sentences of (24) to be parsed.

Originally, **DPOS4C** permitted the attachment of certain subordinating clauses when the word being considered by the parser was subcategorized in the lexicon as a **CS7** subordinating conjunction. **CS7** subordinating conjunctions are words like **WITH** and **WHEN** in (24) that introduce so-called **SVEN** constructions. These latter constructions are characterized as having a **SUBJECT** followed by a participial phrase (**VEN**), such as **WITH UNIT NOT ENGAGED** and **WHEN NEW SAC R[E]C[EI]V[E]D** in (24). However, as **DPOS4C** was originally written in the LSP grammar, only (24a-b) would parse. Sentence (24c) could not parse even though **WITH** in (24b) was subcategorized in the lexicon as a **CS7** word because **DPOS4C** was too constraining. It did not permit the attachment of **SUB7** strings in adjectival fragments, which is how (24c) is analyzed.

The main clause in (24c) consists entirely of the adjectival fragment **UNABLE TO MAINTAIN MINIMUM OIL PRESSURE. DPOS4C** originally stated that **SUB7** strings occur only in strings to the right of the verb or after various object strings such as **OBJECT, PASSOBJ,** and **OBJBE**. The relationship of "right of the verb" or immediately after certain **OBJECT**-strings does not hold in (24c); therefore, we rewrote **DPOS4C** to (25):

(25) DPOS4C = IN CSSTG RE SUB7:
THE PREVIOUS-ELEMENT OF IMMEDIATE SA[15]
IS RV OR OBJECT OF PASSOBJ OR OBJBE OR ASTG.

The Restriction in (25) is characteristic of all D-Restrictions in LSP. Immediately following the name of the Restriction, which is mnemonic except for the required **D** to alert the parser to the type of restriction being fired, is the "housing." The Restriction in (25) is housed in **CSSTG**, and the specific ("RE") expansion under consideration, **SUB7**, follows. Basically, **DPOS4C** requires that the aunt of the present node be a particular syntactic category. As

15. **DPOS4C** requires certain dominance relationships to hold in the tree to obtain a good parse. Therefore, the major stipulation of **DPOS4C** is that the **CSSTG**s under question must occur in certain configurations of nodes in the tree. These considerations are not discussed here. Instead, the reader is directed to Ref. 2 where a full discussion of the various **ROUTINES** that express various relationships and ways of traversing a tree are discussed.

rewritten in (25), **DPOS4C** allows the subordinating clauses under discussion to appear to the right of adjectival clauses while still allowing them to appear to the right of verbs and after **OBJECT**-like strings, as in (24a and b).

No Restrictions had to be tightened further, and only one English Restriction had to be ignored based on Navy CASREP data. This Restriction dealt with loosening the Subject-Verb Agreement Restriction.

English requires plural subjects to co-occur with plural verbs in the present tense, in the present perfect, and with "do" auxiliaries. (The idiosyncracies associated with the verb "be" in English were not altered, because this verb characteristically exhibits its typical behavior in the Navy CASREP domain.) A systematic reflex of this co-occurrence constraint holds between singular subjects and verbs as well. Since CASREPs exhibit ill-formed sentences, it was necessary to ignore this Restriction to allow sentences like (26) to parse with no problems. Ignoring this Restriction allows parsing to proceed in all cases.

(26) INVESTIGATION AND TROUBLESHOOTING OF CAUSE REVEALS FAILURE TO SSDG.

While such a sentence as (26) would be considered ungrammatical from a prescriptive point of view, it is perfectly grammatical, i.e., it must be accounted for by the grammar parsing the messages of which (26) is but one sentence in the corpus. Navy message writers, literate in every respect, may still produce sentences such as (26) when grammatical complexity or external noise diverts their attention from the supposed grammatical rules of "good" English. The parser, therefore, must be capable of recovering from such types of sentences.

5. ADDITION OF NAVY-SPECIFIC RESTRICTIONS

Our grammatical work on developing a sublanguage grammar in a Navy subdomain required that we add 12 "disqualify" or D-Restrictions to the Navy component of the **Restriction Component**. We also added 30[16] "well-formedness" or W-Restrictions. Most of the grammatical work in fine-tuning the LSP grammar and adapting it to the specific Navy subdomain under investigation (namely Navy Starting Air Compressor Casualty Reports) was in the area of adding specific Restrictions to the **Restriction Component** of the grammar. Generally, these additions to the grammar consisted of Restrictions that either optimized the parsing process or constrained the occurrence of certain linguistic strings. A sample of our work in this area follows.

Optimization of Domain-specific Constructions

As stated above, certain Restrictions were written that "optimized" the parsing process. **DNAV1** (in Fig. 3, repeated here as (27) for convenience), is a "disqualify" Restriction.

(27) DNAV1 = IN INTRODUCER RE OPTION LNR:
THERE !S A ":" AHEAD.

DNAV1 ensures that a particular element occurs ahead in the string that is being parsed. By checking ahead of where the parser is at a particular stage of the parsing process, unnecessary time in attachment and backtracking can be avoided. Thus, **DNAV1** ensures that when the **INTRODUCER** string is attached by the parser, further attachment of substrings of **INTRODUCER** will only occur if a colon is somewhere in the remainder of the sentence. This "look ahead" procedure forces the parser to look ahead in the sentence and attach the particular node in question only if certain conditions are met. If no specified string is ahead in the sentence, the node in question is not attached, thereby saving time in attachment and backtracking if necessary. For example, a sentence like (28a) is scanned for the colon at the moment that **INTRODUCER** is attached and its first option **LNR**—a nominal string—is selected.

(28) a. CASREP 002: SAC FAILED.
b. SAC FAILED TO ENGAGE.

16. Numbering discrepancies resulted from subsequent reorderings and removal of rules caused by redundancies. The actual numbers that identify a Restriction are immaterial to their functionality. They are not crucial to the application and therefore were not changed during the last updating.

Since a colon is ahead in (28a), parsing of **INTRODUCER** continues. However, in (28b), because the sentence lacks a colon, **INTRODUCER** does not expand into the **LNR** option. Instead, **NULL** is inserted in the parse tree as a permissible terminal node in the parse tree for the first expansion, and the parser moves on to the next option, attaching **CENTER**, which further expands to the **ASSERTION.** Parsing is correctly concluded. Thus, **DNAV1** was written to ensure that sentences like (28) would parse correctly with minimal backtracking in terms of **INTRODUCER** attachment.

Restricting Occurrences of Domain-specific Constructions

Several Restrictions were written to restrict the occurrences of certain domain-specific constructions. In Section 1, we noted the two kinds of Restrictions: "D" ("Disqualify") and "W" ("Well-formedness"). In Section 1, when describing how an LSP grammar operates and functions to parse sentences, we described one Restriction, **WNAV17**. We do not consider this type of Restriction in any great detail further. Rather, we refer the reader to the discussion of **WNAV17** in Section 1. On the other hand, because of the complexity of a related issue concerning one of the D-Restrictions, we do discuss one of the D-Restrictions in greater detail in the next section.

An interesting case in which the **Restriction Component** had to be modified by the addition of a D-Restriction is a fairly complex one involving the parsing of morphologically identical, or "homophonous," past tense, and past participial parts of verbs. Verbs, such as ***walked*** or ***engaged***, are problematic for deterministic parses in certain environments and constructions, as we see below. We now turn to this problem and a tentative solution obtained in writing a D-Restriction to bar occurrences of a specific construction in our Navy sublanguage grammar.

6. A SUBLANGUAGE PROBLEM REVISITED

The Problem

While English exhibits verbs that have transitive and intransitive uses, it has been argued [9] that verbs in sublanguages do not. This conclusion provides an easier solution to the problem of parsing active sentences and telegraphic passives[17] [10] in sublanguages than the one argued for here. Given this lexical constraint on verbs, confusion in a sublanguage is avoided when intransitive, active, past-tense forms and telegraphic, passive, participial forms of doubly subcategorized verbs[18] are encountered during syntactic parsing.

The way in which parsing of doubly subcategorized verbs was handled in developing a sublanguage grammar for the Navy CASREP domain is made explicit in the discussion below. It is our claim that one solution offered [9] is not a viable one for the sublanguage grammar of this report. The sublanguage investigated here exhibits verbs that are both transitive and intransitive in their usage. Therefore, our lexicon had to incorporate the double subcategorization under question. There simply is no adequate alternative to account for these types of verbs when they are part of the corpus under investigation.

Although the solution offered here is more complex, it is consistent with the sublanguage data processed to date and produces correct and efficient parses. However, it is only a tentative solution given more recent information about the distribution of passive constructions in this sublanguage. Based on some preliminary observations about this new information, we conclude with recommendations for future research.

Discussion

In English (and Navy) passive constructions, the underlying object role of the verb is expressed syntactically as the subject, and the underlying subject role of the verb is optionally expressed syntactically as the object of a prepositional phrase, usually by using the preposition ***by***. The verb also exhibits some morphological change. However, the surface morphology of some verbs is not distinct in the past tense and past participial forms, causing somewhat of

17. Telegraphic passives are passive constructions that do not contain some form of the verb BE, as in the ambiguous sentence *SHIP ATTACKED*, which in one sense is the telegraphic form of *SHIP WAS ATTACKED BY SUBMARINE*.
18. Hereafter, when we refer to "doubly subcategorized verbs," we are referring to verbs that have been subcategorized as transitive and intransitive in the lexicon.

a parsing problem in deterministic, top-down, left to right parsing with backtracking. Characteristically, verbs that are subcategorized for direct objects are transitive and are capable of being passivized, while verbs that do not have direct objects are intransitive and are not capable of being passivized. Thus, (29b) is the passive counterpart of (29a).

(29) a. FCT DISENGAGED THE SAC.
b. THE SAC WAS DISENGAGED BY FCT.
c. ALARM SOUNDED WITH SAC DISENGAGED.
d. DISENGAGED OIL PRESSURE WAS NORMAL.

The subcategorization of transitive verbs leads to the second part of the parsing problem with such parses.

In the LSP grammar, this correspondence between active and passive verbs is expressed by the fact that any verb that is subcategorized for **NSTGO** in its **OBJLIST** (therefore, transitive) in the LSP lexicon is automatically capable of appearing in passive sentences. It can also act as a participial modifier of nouns, as does ***DISENGAGED*** in (29c-d). If the verb is also subcategorized for a **NULLOBJ** in its **OJBLIST** (this captures the grammatical notion of intransitivity), problems arise because several parses are then possible. This situation results from the parser's inability to determine from the form of the word whether the past tense form or the participial form is being used. In some cases, then, bad parses are obtained or, if the sentence is lengthy, the parser can run out of parsing space while it attempts to arrive at correct alternatives.

Such verbs as ***ATTACK*** in (30a-b) will not yield bad parses or run out of parsing space if we assume (like Fitzpatrick [9]) that verbs in a sublanguage are not subcategorized as both transitive and intransitive.

(30) a. AIRCRAFT ATTACKED THREE SHIPS.
b. AIRCRAFT ATTACKED.
c. SHIP ARRIVED FOR REPAIRS.
d. *SHIP ARRIVED NEWPORT FOR REPAIRS.

The English verb ***ARRIVE*** in (30c) is intransitive, as exhibited by the ungrammaticality (*) of (30d). Therefore, verbs like ***ARRIVE*** never take direct objects in English. In a Navy sublanguage, the same facts hold. They do not pose a parsing problem. However, in (30a-b), the English verb ***ATTACK*** exhibits both its transitive (30a) and ambiguously either its intransitive or transitive usage (30b). If it is the transitive usage, then (30b) exhibits the telegraphic passive in which the correct form of the verb ***BE*** is elided. The sublanguage grammar developed for this particular subdomain of Navy messages captures these facts by subcategorizing the verb ***ATTACK*** in the lexicon as having both an **NSTGO** and **NULLOBJ** in its **OBJLIST**. The subcategorization of verbs like ***ATTACK*** that are both transitive and intransitive is a consequence of LSP requiring that subcategorization frames of lexical items be checked and matched with **VERB** and **OBJECT** occurrences in the parse tree. Since all **ASSERTIONs** in the grammar require **VERB**s and **OBJECT**s, and a procedure in the parsing algorithm checks to ensure that verbs co-occur with appropriate object complements, it is necessary that a syntactic category be specified for both transitive and intransitive strings. Thus, if (30a-b) do not co-occur as sentences in the domain under investigation, the verb ***ATTACK*** does not have to be doubly subcategorized. The problems associated with parsing, such as obtaining incorrect or bad parses and "garden-pathing," are avoided. However, since our data did not permit the easier solution, we had to find an alternative for doubly subcategorized verbs.

The sentence ***The horse walked around the barn fell*** is a classic example of "garden-pathing," since the verb ***walked*** can be parsed as a past tense verb or as a participle. For whatever psychological reasons, most native speakers of English will parse the verb ***walked*** initially as the past tense verb, having assigned subjecthood to ***the horse***. This parse will, of course, be altered once the actual main verb of the sentence ***fell*** is reached. The necessary backtracking routine is known as garden-pathing for obvious reasons. Garden-pathing will not occur, on the other hand, in a parallel sentence, such as ***The horse flown around the barn fell*** because the past participial form is morphologically unlike the past tense of the verb ***fly***, namely ***flew***.

As already noted, both the past participial and past tense of the verb ***walk*** are identical in English. No semantic constraints prohibit the lexical items from co-occurring (i.e., ***horses*** can***walk***) and nothing constrains the reader from

initially misunderstanding the sentence as previously described. However, English also permits right-hand participial modifiers of nouns. Additionally, no semantic constraints prohibit the lexical items from co-occurring (i.e., ***horses*** can ***be walked***). When the last word of the sentence under question is reached, however, the reader (or listener) must backtrack and reparse the sentence appropriately.

While this extra effort may create some mild annoyance or amusement for an English speaker, the kind of backtracking that is required to parse such sentences correctly in a deterministic parser such as LSP can be extremely time consuming. If the sentence is sufficiently long or syntactically complex, several adverse consequences must be dealt with. Such backtracking can cause the parser to run out of parsing space and can cause the parser to terminate an incomplete parse because the node limit allocated for work space during parsing has been exceeded. Increasing the node limit is an exercise in futility because one can never be sure if the amount increased will be sufficient to parse the next sentence with a similar problem. Conversely, allowing extremely large work space when it may not be needed is an inefficient use of computer memory and resources.

Another characteristic of the LSP parser can interact unfavorably with doubly subcategorized verbs. Because the LSP parser is a deterministic parser, the order of options in the BNFs can sometimes produce bad parses when words are doubly subcategorized. For example, the sentence ***The SAC engaged with the diesel disengaged*** will obtain a bad parse if the verb ***engaged*** is doubly subcategorized and nouns are permitted as right-hand participial modifiers. In a deterministic parser, the productions that expand the right-hand modifiers of nouns (in this case the subject of *SAC*) will be expanded prior to any consideration of the **VERB** of the **SENTENCE.** Therefore, ***engaged*** will parse as a participial modifier, and when the parser looks at the word ***disengaged***, it will either try to parse it as a right-hand modifier of ***diesel*** if it too has been doubly subcategorized, or eventually parse it as the main **VERB** of the **SENTENCE.** Thus, bad parses will be obtained.

If verbs are not doubly subcategorized [9], problems of garden-pathing, running out of work space in the computer, and obtaining incorrect parses can be avoided. However, the corpus of sentences of a particular domain under investigation must support the claim that sublanguages do not exhibit verbs that are both transitive and intransitive. Even if the domain of application is not all sublanguages (as seems to be implicit in Ref. 9), a weaker version of that claim would still not help us in solving the problem in the domain that we investigated. If the subcategorization problem is domain-specific, certain sublanguages may so constrain their verbs, but this principle does not constrain the verbs in the Navy domain investigated here. We have found two verbs that require **NSTGO** and **NULLOBJ** in their lexical subcategorization frames. These verbs are **INDICATE** and **[DIS]ENGAGE.** Consider, therefore, the verbs and their transitive and intransitive uses in (31).

(31) a. [Srepa 1.1]: NR 4 SSDG STARTED WITH SAC DISENGAGED AND LOW LUBE OIL PRESSURE ALARM *INDICATED*. [intransitive]

b. [Srepb 9.1]: WHILE PREPARING TO CONDUCT PMS CHECK ON MAIN ENGINES, SAC *INDICATED* ZERO LUBE OIL PRESSURE. [transitive]

c. [Testa 23.1]: THE LOW LUBE OIL PRESSURE AND COMPRESSOR FAIL TO *ENGAGE* ALARM ACTIVATED DURING ROUTINE START OF START AIR COMPRESSOR. [intransitive]

d. [Testb 14.2]: STARTING AIR COMPRESSOR *ENGAGED* FOR APPROX TWO MINUTES WHEN LUBE OIL PRESSURE DROPPED BELOW 65 PSI ALARM SETTING. [intransitive]

e. [Testb 14.3]: COMPRESSOR COULD NOT BE *DISENGAGED* FROM EITHER REMOTE OR LOCAL CONTROL LOCATION, FOR APPROX THREE MINUTES FOLLOWING LOW LUBE OIL PRESSURE ALARM. [transitive]

f. [Srepa 1.1]: NR 4 SSDG STARTED WITH SAC *DISENGAGED* AND LOW LUBE OIL PRESSURE ALARM *INDICATED*. [ambiguous][19]

In (31a), we maintain that the verb **INDICATED** is being used in a domain-specific way; in other words, the sentence is to be interpreted something like the following: ***[THE] NR 4 SSDG STARTED WITH [THE] SAC DIS-***

19.The usage of *disengaged* and *indicated* is debatable in (31f). It is included here because of associated problems with parsing that arise if these verbs are doubly subcategorized in the lexicon. While this sentence looks like an apparent counterexample to the theory presented here, it will be shown below that the alternative offered will handle even sentences like (31f) adequately.

ENGAGED, AND [THE] LOW LUBE OIL PRESSURE ALARM [SOUNDED]. Domain-specific knowledge permits us to claim that **ALARM INDICATED** is equivalent syntactically (and semantically) to **ALARM SOUNDED**, which is found elsewhere in the data (32).

(32) [Testa 4.1]: WHILE DIESEL WAS OPERATING WITH SAC DISENGAGED, THE SAC LO ALARM SOUNDED.

In (31a), the verb **INDICATED** is used intransitively, while in (31b), it is used transitively. Given that we do not encode domain or world knowledge in the grammar, the idea that *ALARM INDICATED* and *ALARM SOUNDED* are somehow equivalent is not encoded in the grammar but in another component of the larger expert system that analyzes text.[20] Here we capture the grammatical facts by allowing the verb to be subcategorized for both transitive and intransitive senses to permit parsing to continue in both cases. Similar facts hold for **ENGAGE/DISENGAGE**, which we are assuming to be similar in their syntactic distributions, as in (31c, d, and e). Therefore, the claim that verbs in a sublanguage are not used both transitively and intransitively is not substantiated in this domain, and we must subcategorize verbs as both transitive (**OBJLIST: NSTGO**) and intransitive (**OBJLIST: NULLOBJ**) accordingly. However, we are left with the problem of either garden-pathing or of generating bad parses, and in some cases, running out of work space in the parser for extremely complicated syntactic constructions.

LSP has a mechanism to constrain the occurrences of participial modifiers to the right of nouns. These occurrences are the ones that we felt needed to be constrained because, in parsing, they are the first type of verb-like nodes encountered as the parser traverses from left to right in going from the **SUBJECT** through right modifiers (Right of the Noun) and on to the **VERB**. In the original LSP system as ported from NYU, the occurrence of participial right modifiers was considered **Rare**. Therefore, when the rule for expanding right-hand modifiers of nouns was written, it was written with a **Rare** flag on the option allowing participial modifiers. If the switch was by default left off, then the option was not tried. If the switch was turned on, it permitted the option to be expanded. However, the switch could only be turned off or on at the beginning of the parsing process for each sentence. Changes, therefore, could not be made during the parse. Sentences like (31f) would still be problematic because the switch would have had to be turned on for parsing the noun phrase *SAC DISENGAGED* but turned off for *ALARM INDICATED*.

Therefore, given that a particular verb must be subcategorized for **NSTGO** and **NULLOBJ** because of its occurrences in the corpus and given the type of parser that was used, an alternative solution had to be reached.

A Solution

We attempted to solve the problem of allowing doubly subcategorized verbs in the lexicon while attempting to reduce or eliminate garden-pathing and bad parses as much as possible. To do so, we reanalyzed the corpus of data and noted that the verbs that had to be doubly classified, namely **[DIS]ENGAGE** and **INDICATE**, never occurred as participial right-hand modifiers to nouns when the host nouns were in the **SUBJECT** position of a sentence. We, therefore, decided to constrain participles in this environment.

The grammar has two possible categories for participial right modifiers of nouns, namely **VENPASS** and **ADJINRN** in **RN**. **RN** is the syntactic category for all right-hand modifiers of nouns adjacent to **NVAR** in the parent node **LNR**. **RN** has a sister, **LN**, for the left-hand modifiers of the host noun **NVAR**. **VENPASS** is the syntactic category that captures participial clauses, and **ADJINRN** subsumes adjectives and participles, **LAR**, that are used as adjectives.[21] Figure 6 shows BNFs greatly simplified for expository purposes only.

---

20. We direct the reader to Ref. 3 for a discussion of the *TE*xt *R*eduction *S*yst*E*m. The current report discusses the grammatical work associated with the research and development of **TERSE**.

21. **ADJINRN** also subsumes **LQNR** (right-hand quantifiers of nouns) but this is not of immediate importance here.

```
<SUBJECT> ::=  <NSTG>.
<NSTG>    ::=  <LNR>.
<LNR>     ::=  <LN><NVAR><RN>.
<RN>*R    ::=  <VENPASS> / <ADJINRN>.
<VENPASS> ::=  <LVNR><SA><PASSOBJ>.
<ADJINRN> ::=  <LAR> / <LQNR>.
<LAR>     ::=  <*ADJ> / <VENPASS>.
```

Fig. 6 — Sample BNF rules

We attempted to constrain the distribution of participial constructions in two particular instances in the **SUBJECT**, namely in **VENPASS** and in **ADJINRN** of **RN**. To do this, w wrote a Restriction **DNAV12** (33).

(33) DNAV12 = IN () RN RE VENPASS, ADJINRN:
EITHER IT IS NOT THE CASE THAT ASCEND TO SUBJECT
OR CURRENT WORD IS Q.

Basically, **DNAV12** states that in the iterative or repeatable node **RN** (indicated conventionally by the double parentheses), **RE**garding the options **VENPASS** and **ADJINRN**, it is **EITHER THE CASE THAT** the node **SUBJECT** is not passed through when the routine **ASCEND** is performed in the parse tree, or the **CURRENT WORD** being looked at by the parser is subcategorized as a **Q**uantifier in the lexicon.[22] With Restriction Language syntax and the definition of the **ASCEND** routine aside, **DNAV12** rules out participial constructions in **SUBJECT** position. By incorporating **DNAV12** into the Navy sublanguage grammar, we were able to classify verbs in the lexicon as being transitive, intransitive, or both, as is common in English. Thus, we are able to prevent garden-pathing in similar sentences and to rule out a number of bad parses caused by the interaction of the type of parser used.

Recently, however, one sentence in a later part of the corpus not processed indicates that **DNAV12** is too restrictive or tight. For example, consider (34).

(34) [Srepa 22.4]: SAMPLE DRAWN FROM SAC SUMP, SHOWED 23699 LUBE OIL TO BE CLEAR AND BRIGHT.

We have not processed sentences like (34), but it is clearly a counterexample to our hypothesis that this domain does not permit verbal right-hand modifiers of nouns in **SUBJECT** position. One possible alternative is to allow **DNAV12** to continue to constrain verbals as right-hand modifiers of **SUBJECT** nouns, but further requires that a comma be present in the string ahead as is evident in (34). This solution admittedly is *ad hoc*; however, this is the only sentence in the CASREP documents in our possession that contains such a construction. We would have to investigate additional documents to see if verbal right-hand modifiers exist in **SUBJECT** position but are "cued" by the presence of a comma separating the rest of the **SUBJECT** from the **VERB** of the sentence. Alternatively, as suggested elsewhere (footnote 22), a Restriction could be written in which the characteristic homophony of such forms is

---

22.We will discard the second conjunct of the Restriction here, namely that the options **VENPASS** or **ADJINRN** are permitted if the **CURRENT WORD IS Q**. This stipulation was added when the grammar was fine tuned to account for all of the sentences of the subdomain. Thus, a sentence like (A) mandated the incorporation of the second conjunct in **DNAV12** because in (A), 1/4 is lexically a **Q**uantifier and the expression is embedded in a right-hand adjectival modifier of a host noun **CHIPS.**

(A) [Testa 28.2]: BLADES ARE BENT AND CHIPS, 1/4 INCH DEEP, ARE VISIBLE ON LEADING EDGE.

Furthermore, while it seems rather *ad hoc* that only **Quantifiers** be permitted as right-hand modifiers to nouns in **Subject** position in this domain, it is descriptively adequate. The underlying notion here is that, perhaps for parsing strategies, verbals whose past tense and participial forms are homophonous are not permitted as the right-modifiers of Subjects. This constraint obviously prevents the resulting confusion.

a trigger in the constraint. This alternative, however, is beyond the current capabilities of the parser and must be left for future investigation. Clearly, additional research is required in this area.

REFERENCES

1. E. Marsh and C. Friedman, "Transporting the Linguistic String Project System from a Medical to a Navy Domain," *ACM Trans. on Office Inf. Systems* 3(2), 121-140 (1985).

2. N. Sager, *Natural Language Information Processing: A Computer Grammar of English and Its Applications* (Addison-Wesley, Reading, MA, 1981).

3. K. Wauchope, E. Marsh, and M.K. DiBenigno, "Automated Text Highlighting of Navy Equipment Failure Messages," NRL Report 9154, Nov. 1988.

4. Z. Harris, "Co-occurrence and Transformation in Linguistic Structure," *Language* **33**, 283-340 (1957).

5. D. Perzanowski and B. Potter, "InterFIS: A Natural Language Interface to an Expert System Shell," *Proceedings of the World Conference on Expert Systems*, Jay L. Liebowitz, ed., Dec. 16-19, 1991, Pergamon Press, Elmsford, NY, in press.

6. N. Chomsky, *Aspects of the Theory of Syntax* (MIT Press, Cambridge, MA, 1965).

7. M. Aronoff, *Word Formation in Generative Grammar*, Linguistic Inquiry Monograph 1 (MIT Press, Cambridge, MA, 1976).

8. N. Chomsky, "Remarks on Nominalization," in *Readings in English Transformational Grammar*, R. Jacobs and P. Rosenbaum, eds. (Ginn and Co., Waltham, MA, 1970).

9. E. Fitzpatrick, "Verb Complements in a Deterministic Parser," NRL Technical Memorandum 7590-077, 1983.

10. J. Bachenko, "Verb Transitivity and the Syntax of Telegraphic Passive Sentences," NRL Technical Memorandum 7590-230, 1983.

Appendix A

AN EXAMPLE OF A SENTENCE LOG ENTRY

The following is an example of one of the sentence logs kept during the grammatical work on creating a Navy sublanguage grammar. The entire log is organized numerically according to the actual occurrence of the sentences analyzed in the Navy messages. Sentence IDs appear on the first line with the most recent status of the parse.

In the example below, the word **GOOD** appears after the sentence ID because the grammatical work on this sentence has obtained a good parse. The parsed sentence appears on the next line. On the third line, the date of a particular parsing run appears, as does an acronym of the name of the run. On the next line(s) the intermediate status of the run appears. If the status of the run is good, then a "G" appears. If no parse ("N") or a bad parse ("B") was obtained, then debugging comments appear.

Keeping a sentence log (like the one that follows) ensures that the various reasons that prompted grammatical changes for individual sentences (and grammatical constructions) are part of the project's records for future reference. They are also helpful in tracing and retracing various lines of reasoning for making changes or further altering grammatical changes.

#TESTB 2.1.1. GOOD
LOSS OF LUBE OIL PRESSURE DURING OPERATION NR. 2 SSDG.
4/28/87 run bsa3
G.
02/20/87 run bsa1
B: NPOS and NQ problems.
1. Given amt. of previous work on NQ and ADJINRN and RN1 with NPOS constructions and LISTs, I first looked at LIST N-NPOS, noticing that ptrc.lis allowed FUNC in NPOS on PART in NVAR. This construction was allowed given classes not on LIST. But we can't rule out these classes in these constructions, given PROPULSION GAS TURBINES in testb 15.1.1. Next looked at WNAV19, which deals with ADJINRN constructions.
2. Rewrote WNAV19$LQNR-IN-RN1: forgot to restrict HOST-.
04/11/86 run bid3
G.
11/22/85 run currenttest
G.
11/20,33/85 run trynq and currenttest
N: No parse: suspect DOPT24.
1. Adding Q check for NULLN in NQ to DOPT24STEST
B: NR. 2 wasn't parsing as NQ.
1. changed CORE to HOST wording in DOPT25SNQ01/08/85 run bid4a
G.
10/01/84 run bid1
N: FURTHEST ANALYZED: SSDG
1. adding PARG::= NSTGO
2. adding NVN to .11 of OPERATION; uparrowing; adding Restriction for PARG:DNAV5 (cf. bgramlog)
3. adding NVN to list of attributes and PARG in RNP

Appendix B

AN EXAMPLE OF A GRAMMAR LOG ENTRY

The following is an example of one of the grammar logs kept during the grammatical work on creating a Navy sublanguage grammar. The entire log is organized according to the various sections of the grammar. Therefore, grammatical changes to BNF definitions precede changes to LISTs and both precede changes to Restrictions and other parts of the grammar. The first line of individual entries, as in the example below, identifies the sentence(s) that prompted the specific grammatical change(s) for that entry. This is followed by a brief discussion of the particular part of the sentence that prompted the change, after which the grammatical changes are listed. Earlier versions of rules (if they exist) are then included, so that a history of grammatical changes is recorded. This log helps to keep all grammatical changes and the sentences that prompted those changes in a central location for future reference.

#SREPA 5.1.6

Due to structures like SSDG NR 4 SLIPRINGS and HUB INTERNAL GEAR, NNN was altered and new definitions for LNR1, LN1, and RN1 added. Also added WNAV18 to rule out NULLN and NAMESTG in NVAR of LNR1. Added WNAV19 to restrict RN1 of LNR1 to be NQ or NAV-AREA adjs where host is NAV-PART.

<NNN>::= <LNR1> / <NNN> / <LNR1>.
<LNR1>::= <LN1><NVAR><RN1>.
<LN1>::= <LCDN> / NULL.
<RN1>::= <ADJINRN> / NULL.

Originally:
<NNN>::= <LCDN><*N> / <NNN><LCDN><*N> / -<LCDN><*VING> / -<NNN><LCDN><*VING>.
and LNR1, LN1, RN1 did not exist.

Appendix C

TYPICAL BRANCHING STRUCTURE

LN1 is a low-level left-branching modifier to account for compound Navy nominals. It allows for the following nested construction in compound Navy nominals, as the expression **POTENTIAL OVER TEMP HAZARD** exhibits. In the nested structure shown in Fig. C1, **POTENTIAL** modifies the host noun **HAZARD**, the adverb **OVER** modifies **TEMP[ERATURE]. POTENTIAL** and **OVER TEMP** both modify **HAZARD**.

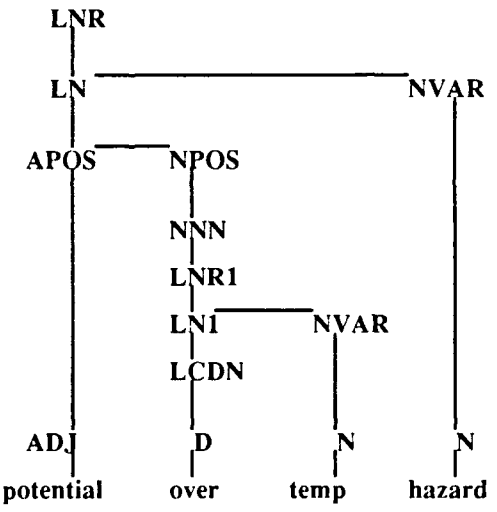


Fig. C1 — Typical nested structure